Bayesian Methods - Ex3b - Parametric Proportional Hazards Model

Jan Vávra

Full assignment in PDF

This assignment is supposed to be solved via JAGS and `library(runjags)`. List through the manual to find what you need.

Data and model description

We will work with the `aids` data from `library(JM)`, see `help(aids)` for more details. It is of both longitudinal and survival data nature. In this part of the exercise, we will focus on the **survival** aspect:

```
set.seed(123456789)
library(JM)
head(aids[,c("patient", "Time", "death", "CD4", "obstime", "drug", "start", "stop", "event")], 15)
```

```
##    patient  Time death       CD4 obstime drug start  stop event
## 1        1 16.97     0 10.677078       0  ddC     0  6.00     0
## 2        1 16.97     0  8.426150       6  ddC     6 12.00     0
## 3        1 16.97     0  9.433981      12  ddC    12 16.97     0
## 4        2 19.00     0  6.324555       0  ddI     0  6.00     0
## 5        2 19.00     0  8.124038       6  ddI     6 12.00     0
## 6        2 19.00     0  4.582576      12  ddI    12 18.00     0
## 7        2 19.00     0  5.000000      18  ddI    18 19.00     0
## 8        3 18.53     1  3.464102       0  ddI     0  2.00     0
## 9        3 18.53     1  3.605551       2  ddI     2  6.00     0
## 10       3 18.53     1  6.164414       6  ddI     6 18.53     1
## 11       4 12.70     0  3.872983       0  ddC     0  2.00     0
## 12       4 12.70     0  4.582576       2  ddC     2  6.00     0
## 13       4 12.70     0  2.645751       6  ddC     6 12.00     0
## 14       4 12.70     0  1.732051      12  ddC    12 12.70     0
## 15       5 15.13     0  7.280110       0  ddI     0  2.00     0
```

It consists of 467 HIV infected patients who were treated with two antiretroviral drugs (`drug`).

Variable `Time` contains the infromation about the time in months from the entrance to the study when either

- patient died (`death = 1`) or
- patient left the study (`death = 0`).

In classical survival model we would denote by

- $T_i \in (0, \infty)$ the time of death of patient i,
- $C_i \in (0, \infty)$ the time of censoring of patient i (leaving the study),
- $X_i = \min\{T_i, C_i\}$ is the observed event time (the one which comes first - death or censoring),
- $\delta_i = \mathbf{1}(T_i <= C_i)$ is the event indicator (1 = death, 0 = censored).

The goal is to model the distribution of the time T_i given covariates. From Censored Data Analysis course we know that neither omitting observations with `death = 0` nor treating all X_i as the true T_i is appropriate. Both approaches would severely underestimate the true survival time.

First, we will eliminate the longitudinal aspect of the data. **Filter the data to one line per patient. Keep the last available row, average the `CD4` cell count across all visits.** $\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ denotes the average value of `CD4`.

```
##    patient  Time death      CD4 obstime drug gender prevOI         AZT start
## 3        1 16.97     0 9.512403      12  ddC   male   AIDS intolerance    12
## 7        2 19.00     0 6.007792      18  ddI   male noAIDS intolerance    18
## 10       3 18.53     1 4.411356       6  ddI female   AIDS intolerance     6
## 14       4 12.70     0 3.208340      12  ddC   male   AIDS     failure    12
## 18       5 15.13     0 7.798241      12  ddI   male   AIDS     failure    12
## 19       6  1.90     1 4.582576       0  ddC female   AIDS     failure     0
##     stop event
## 3  16.97     0
## 7  19.00     0
## 10 18.53     1
## 14 12.70     0
## 18 15.13     0
## 19  1.90     1
```

As it is standard practice in survival models, we will parametrize the distribution of time T_i via its (cumulative) hazard function the same way as in Cox proportional hazards model:

$$h(t) = h_0(t)\,\exp\left\{\mathbf{Z}^\top\boldsymbol{\gamma}\right\}, \qquad H(t) = \int_0^t h(s)\,\mathrm{d}s = \int_0^t h_0(s)\,\mathrm{d}s\,\exp\left\{\mathbf{Z}^\top\boldsymbol{\gamma}\right\} = H_0(t)\,\exp\left\{\mathbf{Z}^\top\boldsymbol{\gamma}\right\}$$

where

- $h_0(t), H_0(t), t \in (0,\infty)$, is the baseline (cumulative) hazard function for individual with $\mathbf{Z} \equiv \mathbf{0}$,
- $\mathbf{Z}$ is the p-dimensional vector of covariates,
- $\boldsymbol{\gamma} \in \mathbb{R}^p$ is the vector of unknown parameters.

In traditional Cox proportional hazards model (function `coxph` from `library(survival)`) the approach is *semi-parametric* - function $h_0(t)$ is assumed to be arbitrary and primary target is on parameters $\boldsymbol{\gamma}$, which is solved by optimizing the partial likelihood.

Here we will assume fully *parametric* form of the hazard function. Consider Weibull distribution with shape parameter $\alpha > 0$ and scale $s > 0$, $\mathsf{Weib}(\alpha, s)$, see `?Weibull`. Then, the corresponding (cumulative) hazard function is a power function:

$$h_{\mathsf{W}}(t) = \frac{\alpha}{s}\left(\frac{t}{s}\right)^{\alpha-1}, \qquad H_{\mathsf{W}}(t) = \left(\frac{t}{s}\right)^{\alpha},$$

when

- $\alpha = 1$, we obtain *constant* hazard function, which corresponds to exponential distribution,
- $\alpha > 1$, we obtain *increasing* hazard function,
- $\alpha < 1$, we obtain *decreasing* hazard function.

Weibull distribution in JAGS is parametrized through parameter $\lambda = s^{-\alpha}$, which yields

$$h_{\mathsf{W}}(t) = \alpha t^{\alpha-1}\,\lambda, \qquad H_{\mathsf{W}}(t) = t^{\alpha}\,\lambda$$

Hence, if we suppose $h_0(t) = \alpha t^{\alpha-1}$, $H_0(t) = t^\alpha$ and $\lambda = \exp\left\{\mathbf{Z}^\top\boldsymbol{\gamma}\right\}$ we get the same expression for the hazard function as in proportional hazards model, now with parametrized baseline hazard function.

Task 1 - Cox Proportional Hazards Model

Fit Cox proportional hazards model (e.g. `coxph` from `library(survival)`) with the following formula:

$$\mathbf{Z_i}^\top \boldsymbol{\gamma} = \gamma_1 D_i + \gamma_2 \overline{Y}_{i\cdot}$$

where

- $\overline{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ denotes the average value of `CD4`,
- D_i is the indicator of `drug` level `ddI`,

Print the summary of the model. Estimate the cumulative distribution function and survival function $S(t) = \exp\{-H(t)\}$ and judge whether the assumption of Weibull distribution may be reasonable.

Task 2 - Bayesian approach - JAGS implementation

Consider the regression formula from Task 1. When the observed event is death `event = 1`, then we observe T_i directly. When the observation is censored, we only know $T_i \in (X_i, \infty)$ and the exact value of T_i is unobserved. We will treat these times as unknown auxiliary variables (*data augmentation*). In JAGS, this is performed by `dinterval`, see details in the manual. It requires proper data preparations.

As discussed above we will assume Weibull distribution for the time to death:

$$T_i \,|\, \alpha, \boldsymbol{\gamma}, \mathbf{Z}_i \sim \mathsf{Weib}\left(\alpha,\, \lambda = \exp\{\mathbf{Z_i}^\top \boldsymbol{\gamma}\}\right)$$

Assume the independent block structure of the prior for model parameters:

$$p(\alpha, \boldsymbol{\gamma}) = p(\alpha) \prod_{j=1}^{2} p(\gamma_j)$$

Choose weakly informative *normal* prior for γ_j and $\alpha \sim \Gamma(1, 0.001)$.

Write down (and print) the model implementation within JAGS.

Task 3 - Running JAGS

Sample two Markov chains using JAGS to approximate the posterior distribution $p(\boldsymbol{\gamma}, \alpha \,|\, \mathsf{data})$. Choose appropriate `burnin` and `thin` by monitoring the trajectories and autocorrelation. Monitor also 3 auxiliary variables for unobserved time of death (arbitrarily chosen).

Warning: working with NA values for right-censoring can cause some unexpected issues. Avoid monitoring "pd", "popt", "ped", monitoring only "deviance", "dic" suffices.

Task 4 - Monte Carlo estimates

Provide summaries including ET and HPD intervals for primary model parameters. Compare them to the outputs in Task1. Estimate the posterior distribution of values of cumulative hazard and survival functions $H(t)$ and $S(t)$ on a dense grid of time values for a patient of reasonable covariate values. Plot the estimated posterior mean and ET credible intervals together with the Breslow estimator from Task~1. Was the choice of parametric family appropriate?